54-55 | 2010

cahiers de praxématique

Corpus, données, modèles

Sophie AZZOPARDI

Felice ADDEO
Hirofumi ANDO
Christophe BENZITOUN
Lolita BERARD
Dominique BOUTET
Alexandra CARIA
Isabel COLÓN DE CARVAJAL
Juliette DALLE
Patrice DALLE
Sophie DALLE-NAZÉBI
Claire DANET
Raphaël DE COURVILLE
Paolo DIANA
Patrick DOAN
Christelle EXARE
Jiayin GAO
Delphine GIULIANI
Clémentine HUGOL-GENTIAL
Simon LANDRON
François LEFEBVRE-ALBARET
Laurence LONGO
Stéphanie LOPEZ
Vassiliki MARKAKI
Damon MAYAFFRE
Roman MILETITCH
Lorenza MONDADA
Samira MOUKRIM
Ahmad NAWAFLEH
Nikola PAILLEREAU
Geneviève PINARD-PREVOST
Catharina PINON
Morgane RÉBULARD
Aude WIRTH
Agnès WITKO

Cahiers de praxématique 54-55, 2010, 233-248

Damon Mayaffre
Laboratoire bases, corpus, langage (B.C.L.) — U.M.R. 7320 (Université Nice-Sophia-Antipolis — C.N.R.S.)

Corpus et web-corpus. Réflexion sur la corporalité numérique

Introduction

Les années 2000 ont vu le triomphe du corpus en linguistique ; non pas que l'objet corpus n'ait existé de longue date auparavant, non pas que les linguistes l'aient ignoré jusqu'alors mais au sens où la linguistique *sans* ou *hors corpus* paraît aujourd'hui une spéculation intellectuelle marginale pratiquée seulement par une minorité. Sémanticiens, phonologues, lexicologues, dialectologues, etc. se revendiquent tous du corpus ; même la syntaxe semble désormais concernée à en croire le dernier numéro de la revue dédiée depuis 2002 entièrement au sujet : *Corpus* [cf. *Corpus* n° 9, 2010, « La syntaxe en corpus » (M. Oliviéri)].

Nous avons déjà par deux fois [Mayaffre 2005 et 2007 a] fait le bilan de la fièvre autour des corpus qui a saisi sinon le monde en tout cas l'hexagone scientifique au début des années 2000. Depuis lors, le mouvement s'est encore conforté : on lira par exemple un bilan documenté dans [Laks 2008], on consultera les actes du colloque thématique du Cercle belge de Linguistique (22-24 mai 2008) [Mellet et Longrée 2009] ; on mentionnera encore le 23e colloque international du Cercle linguistique du Centre et de l'Ouest (université de Poitiers — 5 et 6 juin 2009) intitulé « L'exemple et le corpus : quel statut ? » Et à Lorient, les Journées annuelles de linguistique de corpus (J.L.C.), à côté des Journées internationales biannuelles d'Analyse de données textuelles (J.A.D.T.) dont la prochaine session se tiendra à Liège au printemps 2012, semblent s'être imposées dans le concert national et européen. D'un point de vue théorique encore, on citera le dernier ouvrage de François Rastier : *La mesure et le grain. Sémantique de corpus* (Paris : Champion, 2011).

Dans ce cadre foisonnant, cette contribution proposera deux réflexions.
D'abord, il convient de réfléchir à la concomitance, qui n'a rien d'une coïncidence, du développement de la linguistique de corpus et de la révolution numérique. S'il ne paraît pas faire de doute que le numérique a favorisé l'essor, le partage, l'interrogation des corpus, il faut aujourd'hui faire le point sur ce qu'un corpus numérique (*versus* un corpus papier) représente. Particulièrement, une question reste pendante : le web, qui est le résultat le plus abouti de la révolution numérique en cours, peut-il être considéré comme un corpus ?
Ensuite, si l'on partage le constat de l'inévitabilité du corpus aujourd'hui dans les études en linguistique ou en SHS, il paraît urgent d'approfondir notre réflexion sur la « corporalité » (à l'heure du numérique). Qu'est-ce qui fait qu'un corpus fait corps ? Qu'est-ce qui fait qu'un corpus est un corpus ? En quoi un corpus fait-il sens ? En quoi est-il cohérent et cohésif, homogène, un, unique ? Ici, dans une réduction de la réflexion générale à notre champ particulier de recherche, nous traiterons avant tout des corpus textuels, et par là, nous essayerons de problématiser le parallèle (im)possible entre texte et corpus, entre textualité et corporalité.

1. Corpus papier, révolution numérique, Web-corpus

Il est aujourd'hui admis que le passage du papier au numérique ne représente pas un simple changement technique de support de la culture, de l'information et du savoir humains mais une révolution culturelle, épistémologique, anthropologique aussi, sans guère de précédent dans l'histoire. Dans son dernier ouvrage, l'anthropologue britannique Jack [Goody 2007] s'amuse à ramasser l'histoire universelle en quelques grandes stations que furent (i) l'invention du langage qui nous fit entrer dans l'humanité, (ii) l'invention de l'écrit qui nous fit entrer dans l'histoire (*versus* la préhistoire), (iii) l'invention enfin plus récente mais déterminante de l'imprimerie qui nous fit entrer dans la modernité. Or aujourd'hui, la révolution numérique semble devoir ouvrir un nouveau chapitre de l'histoire humaine et paraît devoir être plus marquante encore que la révolution Gutenberg sur laquelle nous vivons depuis la moitié du XV^e^ siècle : après la modernité, l'hypermodernité s'impose à marche forcée bouleversant notre rapport au

monde, au temps, à l'espace et aux autres ; l'écran remplace le livre ; le texte est devenu hypertexte.

Ici, en ce qui concerne nos seules pratiques scientifiques, et pour s'en tenir, dans le cadre de la présente réflexion, au seul établissement du corpus ou des données, le numérique révolutionne nos points de vue.

Si le numérique dématérialise le texte et délinéarise la lecture [Darnton 2011, Mayaffre 2007-b, Mellet et Longrée 2009, Rastier 2011, Vandendorpe 1999, Viprey 2006, etc. [1]], en amont, il semble avoir eu pour conséquence de matérialiser le corpus ; non pas de le virtualiser donc, mais, paradoxalement, de le matérialiser.

Là où le corpus était seulement perçu, traditionnellement, comme une idée ou une potentialité, le numérique désormais l'incarne, le réifie, le matérialise en le rendant, quelle que soit sa longueur, palpable et manipulable, exploitable et réexploitable, archivable et échangeable. Si le terme de corpus avait peu cours dans les décennies précédentes, avant de s'imposer aujourd'hui, c'est qu'il était sans grande pertinence car sans contrainte voire sans réalité effective ; flasque jusqu'ici, il est devenu désormais un concept dur. Le corpus était en effet un idéal ou une potentialité (« virtuellement tous les textes ou toutes les données susceptibles de m'intéresser ») : c'est aujourd'hui un matériau (« réellement, seuls les textes ou les données que j'ai ***saisis*** en machine et que je peux matériellement soumettre au traitement »). Hier encore horizon (parfois une simple liste de documents éligibles), il est devenu aujourd'hui, à la faveur du numérique, un continent, dont la clôture constitue une limite mais sur lequel il est désormais possible de circuler ou de *surfer*. C'est en ce sens que le développement de la linguistique de corpus (c'est-à-dire le dépassement du mot seul, de la phrase seule ou du texte seul par le corpus, considéré alors comme le macro-objet de la linguistique) est un produit de la révolution numérique. Du *Brown corpus* au *British national corpus* en passant par le *Trésor de la langue française*, des corpus lemmatisés du Lasla aux corpus XMLisés de la *Base de Français Médiéval* en passant par le *Nouveau corpus d'Amsterdam* — sans rien dire des corpus particuliers —, tous les linguistes de corpus

1. Il n'est hélas pas le lieu de revenir ici sur l'élément le plus fort de la révolution numérique : la modification de notre perception du texte et la transformation corrélative de l'acte même de lecture. Ces modifications fondamentales ont été théorisées, décrites et illustrées notamment par les auteurs cités.

travaillent aujourd'hui sur données numérisées, et tous réfléchissent à des méthodes numériques pour traiter leur objet numérique.

Tant et si bien qu'il nous paraît naïf et sans productivité scientifique aujourd'hui de se demander si le web — c'est-à-dire l'ensemble des documents numériques librement consultables de chez soi — peut devenir un corpus. Dans un retournement spectaculaire des choses, la question pertinente n'est-elle pas plutôt de savoir si un corpus peut désormais advenir sans le web et sans le numérique ? Nous retrouvons ici, mais de manière radicale, la réflexion binaire de la littérature anglo-saxonne d'abord développée par [de Schryver 2002 et Flechter 2004], puis par [Hundt, Nesselhauf and Biewer (éd.). 2007] : le *Web as corpus* versus *le Web for building corpus.*

Un corpus peut-il exister aujourd'hui sans le numérique ? Derrière la provocation de cette interrogation — nous n'ignorons pas qu'il y avait des corpus avant la révolution informatique ! —, nous voulons sérieusement poser que la nature numérique des corpus aujourd'hui modifie notre perception de l'objet jusqu'à en amender la définition.

En linguistique (comme ailleurs), les corpus sont avant tout des lieux d'attestation (là où le cerveau est considéré par la linguistique chomskyenne comme un organe d'introspection). Partant, deux objections, simples mais jusqu'ici dirimantes, ont été opposées avec récurrence aux chercheurs sur corpus :

1. Le corpus est-il assez recouvrant, assez représentatif, assez « grand » pour prétendre fournir les attestations espérées ? La linguistique de corpus n'est-elle pas condamnée, en travaillant sur des données empiriques et limitées, à donner des résultats certes avérés mais toujours contestables car fragmentaires ou partiels, infra-significatifs voire anecdotiques ?

2. Dès lors que le corpus serait assez volumineux pour représenter un tout représentatif et une masse critique indiscutable, comment imaginer l'interroger dans son immensité ? Comme les données sont seulement ce que l'on se donne, un corpus trouve sa raison d'être seulement dans notre capacité à l'interroger : c'est un objet heuristique et rien d'autre que cela [*Corpus* 2002]. En d'autres termes : de quoi un (grand) corpus, si possible exhaustif, peut-il bien attester sans moyen de description, de navigation, d'interrogation suffisamment puissants ?

À ces deux niveaux de questionnement, le numérique apporte des réponses importantes à défaut d'être définitives.
Par leurs tailles nouvelles (parfois plusieurs milliards de mots) et par les outils d'investigation toujours plus perfectionnés alliant approche qualitative et approche quantitative, les corpus numériques proposent de nouveaux observables et de nouvelles données en linguistique jusqu'à modifier notre rapport à l'empirie du langage. Il s'agit là de la démonstration centrale de l'ouvrage récent de François Rastier.

> La constitution et l'analyse de corpus sont en passe de modifier les pratiques voire les théories en lettres et sciences sociales. Toutes les disciplines ont maintenant affaire à des documents numériques et cela engage pour elles un nouveau rapport à l'empirique.
>
> (Rastier 2011 : 12[1])

Nous illustrerons ce propos par l'entreprise *Google Books*, par l'outil en ligne mis au point par les ingénieurs de Google en décembre 2010 (Books Ngrams Viewer : `http://ngrams.googlelabs.com/`), et par le logiciel *Hyperbase* (U.M.R. 6039, *Bases, Corpus, Langage*, Université de Nice — C.N.R.S.) qui dans sa version 2011 s'articule sur les milliards de sorties *Google Books* pour en faire un traitement logométrique perfectionné tel que l'A.D.T. et la statistique textuelle l'ont développé ces dernières décennies [Brunet 2012].
Illustration — Si l'historien américain du livre et de la lecture Robert [Darnton 2011] regrette légitimement que l'entreprise *Google Books* soit une entreprise privée et à but lucratif, il reconnaît qu'il s'agit là d'une initiative sans précédent dans l'histoire culturelle mondiale. Après avoir signé, en 2006, une convention avec cinq grandes bibliothèques (New-York, Universités de Harvard, du Michigan et de Stanford et la bodleienne d'Oxford), puis avec une quarantaine de bibliothèques les années suivantes, *Google Books* se propose de scanner tous les livres de la planète[2]. Ainsi, en 2010, après seulement 4 ans de travail, *Google Books* pouvait communiquer avoir déjà saisi 4% du fonds mondial disponible, soit 15 millions de livres, 500 milliards de

1. Dans le détail, on notera dans ce passage que l'auteur réfléchit dans un même mouvement à « l'analyse de corpus » et au travail sur « documents numériques », comme si la première notion recouvrait nécessairement la seconde.

2. Cf. « Quantitative Analysis of Culture Using Millions of Digitized Books Sciencexpress », in *SciencExpress*, 16 décembre 2010 ou les informations données en ligne par *Google Books* dans la rubrique « about Google Books ».

mots (dont 44 milliards pour le français). D'ici une dizaine d'années, ce sera le quart du patrimoine livresque mondial qui sera accessible ; à terme, c'est l'ensemble des livres, documents et archives qui est visé [1]. Bref, *Google Books* constitue d'ores et déjà, la plus grande bibliothèque/librairie ayant jamais existé ; bibliothèque universelle que les trois milliards d'internautes planétaires peuvent instantanément consulter gratuitement ou par paiement [2], de leur laboratoire, de leur foyer ou de leur téléphone portable.

Pour explorer son fonds, *Google Books* se dote d'outils de plus en plus performants. Outre les traditionnels algorithmes de recherche ultra-puissants et ultra-rapides (quelques millièmes de secondes pour traiter des millions de documents), outre les outils de navigation dont on ne présente plus les performances, *Google Labs* offre, en ligne, des outils fréquentiels qui permettent d'étudier la distribution chronologique de grams simples (un mots), de bi-grams (deux mots) ou des n-grams (n mots) des années 1 500 à 2 000 (illustration 1).

Forts de ces données fréquentielles sommaires — il s'agit de simples fréquences relatives — mais gigantesques (le calcul s'opère sur environ 43 milliards de mots), Étienne Brunet et le logiciel Hyperbase permettent depuis le printemps 2011 un traitement qui met en œuvre le meilleur de la statistique textuelle [Brunet 2011]. Par exemple, l'analyse factorielle des correspondances (A.F.C.) peut être mobilisée dans le logiciel, par simple clic, pour décrire l'usage des adverbes temporels au fil des siècles et des décennies (figure 2).

Bien sûr, on prendra soin d'interpréter les figures 1 et 2 avec la plus grande prudence. Les objections à ce type d'analyse sont évidentes et renvoient au sujet de cette contribution : les données de *Google books* sont incertaines au sens où les chercheurs peuvent difficilement connaître leur origine. *Google books* n'est pas un corpus, mais un corps sans (en-)têtes qui rassemble aveuglément, loin de toute hypothèse de travail, des dizaines de milliers de livres tous genres confondus.

1. Nous avons fait ailleurs un bref bilan de la numérisation des archives françaises. [Mayaffre D., 2012, à paraître, « Les mots, le texte, les corpus et l'archive : l'historien face au linguistique. Logométrie et analyse du discours »].

2. La critique la plus forte contre *Google Books* est la fausse gratuité, puisque l'accès aux livres dans leur intégralité est souvent payant. Nous formulerons nous-même ce reproche, non sans avoir rappelé que l'accès à toute bibliothèque (sans parler de l'achat des livres en librairie) a toujours été payant par abonnement, carte d'entrée, prêts des livres.

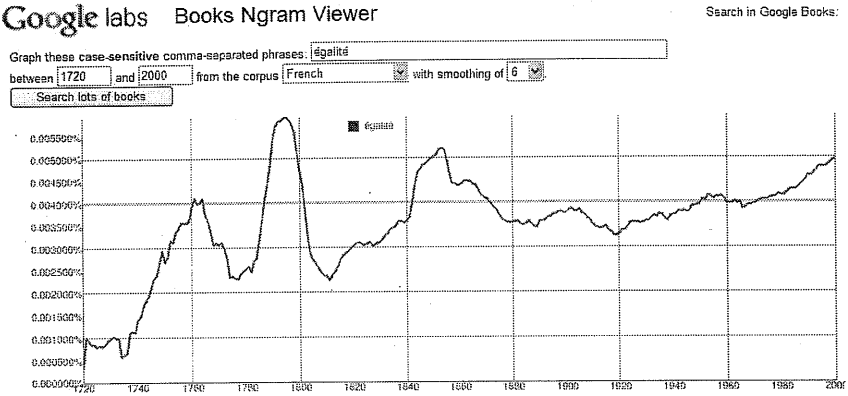

Figure 1. — Fréquence relative de « égalité » dans *Google books* (1720-2000).

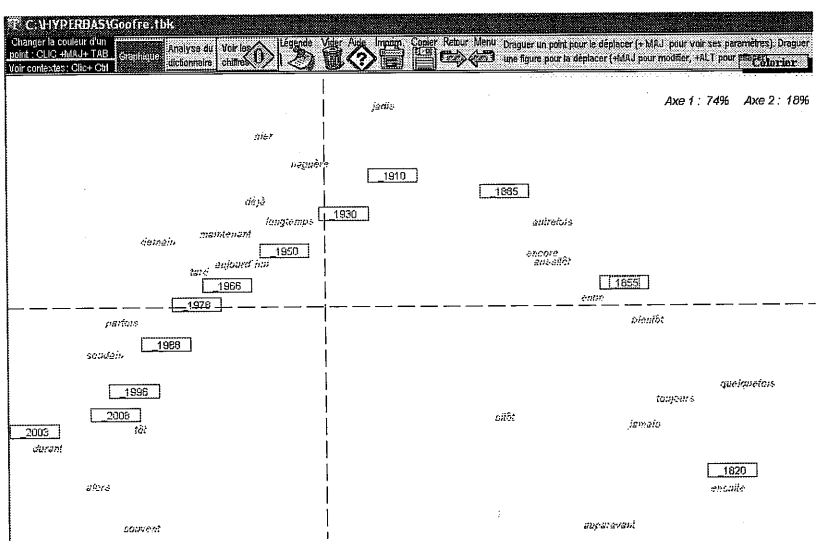

Figure 2. — A.F.C. des adverbes temporels dans *Google Books* (1820-2000). Sortie machine d'Hyperbase.

Pourtant, devant la masse des livres concernés (ici 43 milliards de mots français), et, un jour, devant leur quasi-exhaustivité, ce type de traitement témoigne de quelque chose (de quoi ?) et participe, de fait, à l'heuristique des sciences du langage. Respectivement ici, peut-être peut-on conclure que « l'égalité » (figure 1) ne fut jamais autant réclamée que durant les périodes révolutionnaires qui entourent 1789 d'abord, puis 1848 ; ou encore remarquer que l'explosion fréquentielle « d'égalité » à la fin du XVIII^e^ siècle est précédée d'une période où les revendications égalitaires semblent avoir été mises sous l'éteignoir. Quant aux marques temporelles (figure 2), elles évoluent continûment dans une chronologie impeccable entre 1820 et 2000 — preuve qu'il existe une logique ou un ordre (ici l'axe historique) dans les données linguistiques étudiées [1] ; et on réfléchira, par exemple, au sur-usage rhétorique au début du XIX^e^ siècle de la paire « toujours »/« jamais », auquel répond aujourd'hui une approche plus nuancée avec « souvent ». De la même manière, par économie syllabique sans doute, « sitôt » ou « quelquefois » prisés au début de la période semblent disparaître au profit de « tôt » ou de « parfois ».

De tout cela, on retiendra que les données numériques de grande ampleur, et leur traitement (ici quantitatif) systématique donnent à voir des choses que l'appréhension humaine pouvait certes pressentir mais difficilement attester. Avec les grands corpus numériques, c'est à la fois notre rapport à l'empirie du langage qui est modifié et une nouvelle heuristique des phénomènes langagiers ou discursifs qui peut être envisagée.

2. La corporalité : le corpus comme un texte ?

De la lettre au corpus, en passant par le mot ou la phrase, et en s'arrêtant sur le texte, la linguistique de corpus telle que l'entendent par exemple [Aijmer and Altenberg (éd.) 2002 ; Biber, Conrad & Reppen 1998 ; Habert, Nazarenko et Salem 1997 ; Sinclair 1991 ; Tognini-Bonelli 2001 ; Rastier 2011 ; Williams 2005, etc. [2]] procède

1. Les praticiens de l'A.D.T. auront remarqué la forme parabolique particulière de l'A.F.C., connue sous le nom d'« effet Guttman ».

2. Il va de soi que ces auteurs, entre eux, divergent sur certains points. Nous les rassemblons ici en tant que linguistes ayant pris à bras le corps l'objet corpus et

à/de l'extension de l'objet de la linguistique vers des réalités ou des globalités toujours plus vastes et toujours plus complexes (*grosso modo* et sans entrer dans aucun modèle linguistique : lettres < syllabes < mots < syntagmes < phrases < paragraphes < chapitres < textes < corpus textuels).

Par facilité sans doute, à chaque palier de complexité franchi, l'analyste a eu tendance à se retourner vers le palier inférieur pour en faire remonter des schémas d'analyse qui lui étaient familiers. Ainsi, hier, par exemple, alors même qu'on passait du phrastique au transphrastique, s'est-on imaginé établir — en vain — une *grammaire du texte* comme il en existait une de la phrase.

Ainsi, aujourd'hui, en passant du texte au corpus peut-on envisager expliquer — avec fruit — la « corporalité » comme on explique la textualité ; et précisons que la tâche s'annonce passionnante mais d'autant plus compliquée que la notion de textualité est elle-même à peine stabilisée en linguistique textuelle et objet encore de riches discussions. Peut-on considérer un corpus comme un texte ? Peut-on considérer un corpus textuel comme un macro-texte qu'il s'agirait alors de traiter avec des outils théoriques en partie balisés par Hjelmslev ou Bakhtine, Hasan, Halliday, Adam ou Rastier ?

Ce sont, au fond, les questions qu'avec la plupart des auteurs contemporains l'on est en droit de se poser. On se gardera bien d'apporter une réponse définitive à une interrogation qui engage sinon l'avenir de la linguistique de corpus en tout cas son intérêt actuel. Mais plusieurs indices, comme autant de pistes de réflexion, peuvent être pressentis. Deux méritent d'être rappelés ; nous verrons qu'ils sont nuancés ; et que dans ces nuances réside un programme de recherche.

Cohérence-cohésion du texte/cohérence-cohésion du corpus

Si l'on peut supposer que le corpus, à l'image du texte, est un ensemble cohérent et cohésif, il l'est nécessairement de manière différente. En s'aventurant dans le distinguo heideggerien sans doute peut-on prétendre que la cohésion-cohérence d'un texte lui est ontologique ; la cohésion-cohérence du corpus lui est ontique. Le texte est cohérent

l'idée d'une linguistique de corpus. Quelques grands ancêtres pourraient être ajoutés comme Firth ou Palmer.

par nature, par essence, par définition [1] ; le corpus l'est par existence — en tant qu'*étant* —, par construction, par hypothèse.

Il ne s'agit pas ici de naturaliser (ontologiser) l'objet texte qui est lui-même un objet construit, artefactuel comme nous l'avons clairement souligné, après d'autres, dans [Mayaffre 2007-b], mais de rappeler que le texte existe — sous une forme ou une autre — dans la société sans l'analyse scientifique, là où le corpus existe uniquement en laboratoire par le fait du chercheur, et seulement le temps de la recherche. Le texte est un construit, mais un construit social ou culturel « de première main » par le fait du couple auteur-lecteur (et par l'intermédiaire d'un éditeur). La construction du corpus textuel est, elle, de « seconde main », toujours *ad hoc*, par le seul jeu de l'analyste.

La textualité — ce qui fait qu'un texte est un texte — est définitoire du texte et peut être perçue par tout lecteur, *sans quoi il n'admettrait pas qu'il s'agit là d'un texte*. La corporalité, elle, est une pétition, un espoir, un postulat, le fruit d'un travail singulier ou d'une projection particulière évidente pour le seul chercheur. Exprimée en langage mathématique, la cohésion-cohérence d'un texte est axiomatique ; la cohésion-cohérence du corpus est hypothétique. Bref, un texte qui ne serait pas cohérent-cohésif ne serait plus un texte. Un corpus qui n'est pas cohérent-cohésif est seulement un corpus manquant de pertinence et sans doute d'efficacité heuristique [2].

La différence est donc importante. Pourtant, elle ne nous paraît pas définitive. Quoique d'une autre nature, la cohérence-cohésion du corpus est l'enjeu de la linguistique de corpus exactement comme la cohérence-cohésion du texte est celui de la linguistique textuelle : c'est cette tension commune vers une textualité/corporalité, conçue avec [Charolles 1995 : 10] comme « principe général gouvernant l'interprétation » du texte/corpus, qui rapproche les deux disciplines. Simplement, si du point de vue de la cohérence-cohésion du texte, de [Halliday et Hasan 1976] à [Calas (dir.) 2006] ou [Adam 2008], l'essentiel est déjà réalisé, du point de vue du corpus, l'essentiel reste à faire, même si les travaux de [Viprey nt. 2006] sur la micro-distribution des unités et la

1. Les « propos incohérents » qu'essayent de tenir certains auteurs n'en peuvent mais. Et c'est alors, précisément, cette incohérence du propos qui fait la cohérence du texte.

2. Le lecteur aura remarqué le parti pris de mentionner ensemble, globalement, la *cohérence* et la *cohésion*. Dans le détail, et de manière hiérarchique, il serait facile de montrer que la *cohésion* du corpus pose plus de problème encore que sa *cohérence*.

texture balisent une partie du terrain. Et à ce stade, pressentons seulement, d'un point de vue méthodologique, que le parallèle texte/corpus et textualité/corporalité demandera quelques ajustements : si le texte et la textualité peuvent encore être considérés comme des objets micro réclamant l'approche qualitative, le corpus et la corporalité, en tant qu'objets macro, semblent exiger une approche aussi quantitative.

Sérialité du corpus/linéarité du texte

Si texte et corpus (textuel) présentent certaines similarités au point que l'on peut envisager entre eux un simple rapport d'échelle, une différence profonde de structure semble les distinguer : le corpus est fondamentalement un objet *sériel* [Mayaffre 2002], le texte, lui, est d'abord un objet *linéaire*.

Le corpus est une *collection* de textes réunis sur la base d'hypothèses de travail. Au-delà du stade critique d'une collection de textes qui en compterait un seul, les corpus peuvent donc être considérés comme des séries. (Et faut-il souligner encore ici que les séries, en linguistique comme ailleurs, se prêtent bien au traitement statistique ?)

Certes, certaines séries sont ordonnées linéairement. Nous pensons aux *séries textuelles chronologiques* dont André Salem a décrit les caractéristiques [Habert, Nazarenko, Salem 1997 : 207 et *sq.*] et qui par leur *progression* peuvent à juste droit être traitées comme des textes : il s'agit-là d'un champ de recherche à part entière auquel il conviendrait de s'affronter.

Mais hors des séries textuelles chronologiques, la plupart des corpus-séries n'ont pas de structure linéaire évidente ; de manière significative leurs parties (les textes qui les composent ou des regroupements de textes que l'on aura constitués en parties) peuvent être indifféremment ordonnées sans que le traitement en soit changé. Ainsi, par exemple, dans le corpus de la campagne électorale 2007, les textes des candidats Laguiller, Buffet, Royal, Bayrou, Sarkozy et Le Pen, etc. peuvent contraster et se singulariser indépendamment de leur ordre de saisie en machine et de leur ordre de traitement.

Si le corpus est avant tout sériel donc, et non nécessairement organisé linéairement, le texte lui est toujours linéaire. À l'exception de quelques productions surréalistes ou de quelques jeux d'auteurs

marginaux en littérature [1], un texte a toujours un commencement, un prolongement et une fin ; il peut être défini comme une suite [Maingueneau 1996 : 81 ou Détrie, Siblot, Vérine 2001 : 349] ; et l'élément fondamental de sa lecture est sa progression, pour nous de gauche à droite, de haut en bas. Contrairement aux parties du corpus, les parties du texte (ses phrases, ses chapitres, ses séquences...) ne peuvent être inversées sans remettre en cause l'édifice. Certes, depuis l'abandon du rouleau pour le codex ou le *polyptychon*, rien n'interdit au lecteur de briser par sa lecture cette progression implacable et de papillonner aléatoirement d'une page à l'autre, d'arrière en avant, ou de chapitre en chapitre. Certes encore, aujourd'hui, le numérique permet des lectures hypertextuelles dont la caractéristique est justement de s'affranchir du linéaire : c'est cette révolution majeure évoquée dans une note *supra* et sur laquelle nous revenons immédiatement *infra*. Mais il n'en reste pas moins vrai que la linéarité apparaît irréductible à la textualité et constitue le socle de sa définition [2].

Suite continue *versus* série discontinue, linéarité du texte *versus* sérialité du corpus : touchons-nous donc cette fois-ci à une différence définitive ? Non pourtant.

Objet linéaire, *d'abord*, le texte est aussi traversé de sérialité et de réticularité : c'est l'apport essentiel des riches travaux A.D.T. depuis plusieurs années (cf. notamment la dizaine de volumes des actes des J.A.D.T. depuis 1990). Particulièrement, les travaux sur les co-occurrences attestent de corrélations sémantiques ou d'échos isotopiques qui traversent le texte sans être soumis au seul ordre linéaire de celui-ci [Viprey 2006 ; Mayaffre 2008].

Objet sériel, *en premier*, le corpus est — ne serait-ce que par ce qu'il est composé de textes linéaires — traversé par la linéarité et la séquentialité : c'est l'apport essentiel des travaux récents de [Brunet 2007 et 2008] ou de [Mellet et Longrée 2009] qui mettent à jour l'organisation topologique et linéaire de gros corpus.

1. Précisément il s'agit là de jeux, dont la règle sous-jacente est bien la linéarité attendue... et transgressée.

2. Citons ici la concession définitive du linguiste qui a remis le mieux en cause cette linéarité pour introduire dans le traitement la réticularité : « Nul ne saurait mettre en doute qu'un texte se manifeste dans l'ordre du temps et/ou de l'espace orientés, se caractérise par un début, un milieu, une fin, ordonnés et non interchangeables, et ce à quelque échelle que ce soit de l'organisation macro-séquentielle à la fine succession des périodes » (Viprey 2006 : 74).

Autrement dit, se dessine un double mouvement scientifique qui venant de deux pôles opposés converge en un programme de recherche commun : la linguistique de corpus ou l'A.D.T., partant de la série, vise aujourd'hui à réintroduire la linéarité dans ses traitements. La linguistique textuelle, partant du linéaire, admet la sérialité comme élément complémentaire de son objet. À la suite de Jean-Michel Adam qui, venant du texte, a présenté l'unité de ce programme aux J.A.D.T. 2006 [Adam 2006] avant d'en esquisser des pistes dans [Adam 2008 : 179-182], nous avons essayé, venant du corpus, d'en souligner quelques enjeux [Mayaffre 2007 c].

Dit en des termes admis depuis longtemps en linguistique, il s'agit de rappeler que toute écriture/lecture (celle d'un texte ou celle d'un corpus ; *a fortiori* celle d'un *corpus textuel*) articule une compétence syntagmatique et une compétence paradigmatique. Longtemps essentiellement paradigmatique, la linguistique de corpus qui se développe aujourd'hui à la faveur du numérique, et l'approche statistique des données textuelles qui bénéficie d'un programme riche de plusieurs décennies doivent prendre en compte désormais aussi la dimension syntagmatique, séquentielle et plus généralement encore co(n)textuelle de son objet. Ceci constitue la condition pour forger la « corporalité » dans son entier et espérer aboutir à une sémantique de corpus accomplie.

Références bibliographiques

ADAM J.-M., 2006, « Autour du concept de texte. Pour un dialogue des disciplines de l'analyse de données textuelles », Conférence d'ouverture aux J.A.D.T. 2006. En ligne sur Lexicométrica (`www.cavi.univ-paris3.fr/lexicometrica/jadt/JADT2006-PLENIERE/JADT2006_JMA.pdf`), consulté le 15 mai 2011.

ADAM J.-M., 2008 (éd. revue et augmentée), *La linguistique textuelle. Introduction à l'analyse textuelle des discours*, Paris, Colin.

AIJMER B. & ALTENBERG K. (éd.), 2002, *Advances in Corpus in Corpus Linguistics*, Amsterdam, Rodopi.

BIBER D., CONRAD S. & REPPEN R., 1998, *Corpus Linguistics. Investigating Language, Structure and Use*, Cambridge, Cambridge University Press.

BRUNET E., 2007, « Fréquences et séquences. Mise en œuvre dans Hyperbase », *Lexicométrica* numéro spécial (consulté le 15 mai 2011). Web : http://lexicometrica.univ-paris3.fr/numspeciaux/special9/brunet.pdf.

BRUNET E., 2008, « Les séquences (suite) », in HEIDEN S. & PINCEMIN B. (éd.), J.A.D.T. 2008. Lyon, PUL, 253-266.

BRUNET E., 2011, *Ce qui compte. Écrits choisis*, Paris, Champion.

BRUNET E., 2012, « Au fond du GOOFRE, un gisement de 44 milliards de mots », in *J.A.D.T. 2012, Proceedings of the 11th International Conference on Textual Data Statistical Analysis*, in DISTER A., LONGRÉE D. & PURNELLE G. (éd.), Bruxelles, université de Liège/facultés universitaires Saint-Louis, 2012, 8-21.

CALAS F., 2006, *Cohérence et discours*, Paris, PUPS.

CHAROLLES M., 1995, « Cohésion, cohérence et pertinence du discours », *Travaux de linguistique* 29, 125-151.

CORPUS, 2002-2011, http://corpus.revues.org/.

DARNTON R., 2011, *Apologie du livre*, Paris, Gallimard.

DÉTRIE C., SIBLOT P. & VERINE B. (dir.), 2001, *Termes et concepts pour l'analyse du discours. Une approche praxématique*, Paris, Champion.

FLETCHER W. H., 2004, « Making the web more useful as a source for linguistic corpora », in CONNOR U. & UPTON T. (éd.), *Corpus Linguistics in North America 2002. Selections from the Fourth North American Symposium of the American Association for Applied Corpus Linguistics*, Amsterdam, Rodopi.

GOODY J., 2007, *Pouvoirs et savoirs de l'écrit*, Paris, La dispute.

HABERT B., NAZARENKO A. & SALEM A., 1997, *Les linguistiques de corpus*, Paris, Colin.

HALLIDAY M. A. K. & HASAN R., 1976, *Cohesion in English*, Londres, Longman.

HUNDT M. NESSELHAUF N. & BIEWER C. (éd.), 2007, *Corpus Linguistics and the Web*, Amsterdam & New York, Rodopi.

LAKS B., 2008, « Pour une phonologie de corpus », *Journal of French Language Studies*, 18, 3-32.

MAINGUENEAU D., 1996, *Les termes clés de l'analyse du discours*, Paris, Seuil.

MAYAFFRE D., 2002, « Les corpus réflexifs : entre architextualité et hypertextualité », *Corpus*, 1, 2002, 51-69.

MAYAFFRE D., 2005, « Rôle et place du corpus en linguistique. Réflexions introductives », in VERGELY P. (éd.), *Actes du colloque JETOU'2005*, Toulouse, Université de Toulouse-Le-Mirail, 5-17.

MAYAFFRE D., 2007 a, « Effervescence autour des corpus », in BALLARD M. & PINEIRA C. (dir.), *Corpus en linguistique et en traductologie*, Arras, Artois Presses Université, 61-71.

MAYAFFRE D., 2007 b, « Philologie et/ou herméneutique numérique : nouveaux concepts pour de nouvelles pratiques », in RASTIER F. & BALLABRIGA M. (éd.), *Corpus en Lettres et Sciences sociales. Des documents numériques à l'interprétation*, Toulouse, PUT, 15-26.

MAYAFFRE D., 2007 c, « L'analyse de données textuelles aujourd'hui : du corpus comme une urne, au corpus comme un plan. Bilan sur les travaux actuels de topographie/topologie textuelle », *Lexicométrica*, n° spécial (consulté le 15 mai 2011). Web : `http://lexicometrica.univ-paris3.fr/numspeciaux/special9/mayaffre.pdf`.

MAYAFFRE D., 2008, « De l'occurrence à l'isotopie. Les co-occurrences en lexicométrie », *Sémantique & Syntaxe*, 9, 53-72.

MELLET S. & LONGRÉE, D., 2009, *New Approaches in Text Linguistics*, Amsterdam, John Benjamins.

RASTIER F., 2011, *La mesure et le grain. Sémantique de corpus*, Paris, Champion.

SCHRYVER G.-M., 2002, « Web for/as corpus : a perspective for the African languages ». *Nordic Journal of African Studies*, 11 (2), 266-282. Consulté le 16 mai 2011 : `http://tshwanedje.com/publications/webtocorpus.pdf`.

SINCLAIR J. M., 1991, *Corpus, Concordance, Collocation*, Oxford, Oxford University Press.

TOGNINI-BONELLI E., 2001, *Corpus Linguistics at Work*, Amsterdam, John Benjamin's Publishing.

VANDENDORPE Ch.,
1999, *Du papyrus à l'hypertexte. Essai sur les mutations du texte et de la lecture*, Paris, La Découverte.

VIPREY J.-M., 2006, « Structure non-séquentielle des textes », *Langages*, 163, 71-85.

WILLIAMS G. (éd.),
2005, *La linguistique de corpus*, Rennes, PUR.